



Early Journal Content on JSTOR, Free to Anyone in the World

This article is one of nearly 500,000 scholarly works digitized and made freely available to everyone in the world by JSTOR.

Known as the Early Journal Content, this set of works include research articles, news, letters, and other writings published in more than 200 of the oldest leading academic journals. The works date from the mid-seventeenth to the early twentieth centuries.

We encourage people to read and share the Early Journal Content openly and to tell others that this resource exists. People may post this content online or redistribute in any way for non-commercial purposes.

Read more about Early Journal Content at <http://about.jstor.org/participate-jstor/individuals/early-journal-content>.

JSTOR is a digital library of academic journals, books, and primary source objects. JSTOR helps people discover, use, and build upon a wide range of content through a powerful research and teaching platform, and preserves this content for future generations. JSTOR is part of ITHAKA, a not-for-profit organization that also includes Ithaka S+R and Portico. For more information about JSTOR, please contact support@jstor.org.

THE PROBABILITY CURVE¹

BY WILLIAM L. HART²

In many parts of the general theory of statistics and in innumerable particular problems of a statistical nature, fundamental use is made of the normal probability curve. There is great danger in using any tool if its exact field of applicability is unknown and if its internal mechanism is not understood. A danger of this type is in view whenever a statistical worker uses the probability curve. For, of all the tools furnished by pure mathematics to the person applying mathematical methods, the probability curve, or law of errors, is one of the least perfect from a logical point of view. It is not the purpose of this paper to present a cure for these logical deficiencies or to give a mathematical discussion of the derivation of the probability curve. The reader who is interested in results of this character should consult the references listed in the footnotes of this paper.

In Section 1 of the present discussion, an example will be given showing how the normal probability curve arises naturally in certain statistical problems. Another example of a type apparently similar to the previous one will then be given where, nevertheless, the same procedure as before does not lead to a normal curve. In spite of the fact that the normal curve has a field of applicability the extent of which has not been clearly defined, there are certain conditions, met many times in practical work, under which the normal law of probability holds. In Section 2 one set of such conditions will be given. In Section 3 certain useful properties of the normal curve will be described.

1. Statistical examples and frequency curves. Let us consider the analysis of the results which, under ideal conditions, would be obtained in firing a gun at a target T in a plane. In figure 1 let T represent the target and suppose that the gun is in the direction

¹ Read before the Minnesota Section on May 3, 1920. Discussion is desired and should be sent to the Editor.

² Associate Professor of Mathematics, University of Minnesota, Minneapolis, Minn.

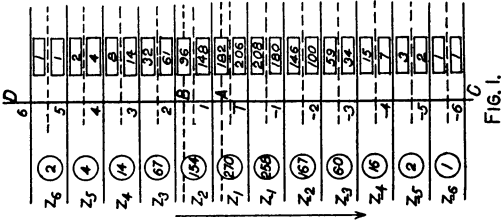


FIG. 1.

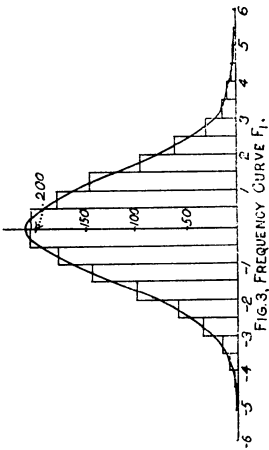


FIG. 3, FREQUENCY CURVE F_1 .

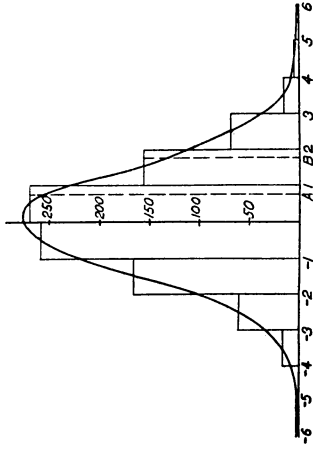


FIG. 2, FREQUENCY CURVE F_2 .

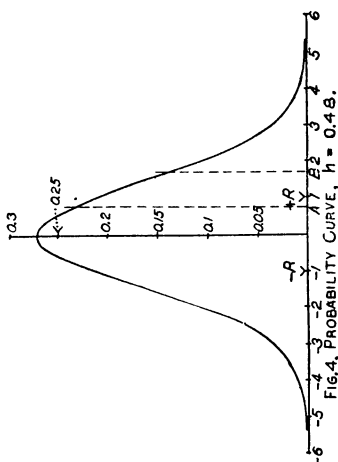


FIG. 4, PROBABILITY CURVE, $h = 0.48$.

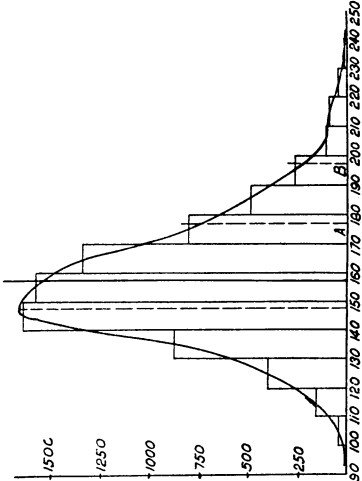


FIG. 5, FREQUENCY CURVE.

EXAMPLES OF FREQUENCY AND PROBABILITY CURVES

indicated by the arrow. Draw a line CD through T and the gun and lay off a scale on CD as indicated in the figure. The full lines through the division points of the scale divide the plane into zones Z_1, Z_2 , etc., and Z_{-1}, Z_{-2} , etc. Suppose, now, that a great many shots are fired at T with a perfect gun, perfectly aimed, on a day when all weather conditions are perfectly normal. Under such conditions, every shot should hit T provided the powder charge used in firing had the proper weight and the projectile fired had been made with absolute accuracy. Let us assume that these last two ideal conditions are not satisfied. Consequently, the shots fired at T will fall in varying numbers in the different zones. Assume that the results of the firing are as shown by the numbers in the circles in figure 1, where each number indicates how many projectiles of the total fell in the corresponding zone.

Consider the construction of a graph of the results as given in figure 2. The horizontal scale is a duplicate of that on CD . The vertical scale has been chosen at will. On each section of the horizontal scale erect a rectangle whose height equals the number of projectiles which fell in the zone corresponding to the section. The resulting figure is called a histogram.³ In the units used for the scales in figure 2, the total area of the histogram equals the number of shots fired, multiplied by the width of a zone, which in this case is 1. If we were to fire a single additional shot, then, on the basis of our experience as given by figure 2, the probability of the projectile falling in the part of the plane between lines in figure 1 through A and B is approximately equal to *the area of the histogram between vertical lines through A and B in figure 2 divided by the total area of the histogram*. Let us call this characteristic property of the histogram the *probability property*.

In figure 2 mark the centers of the tops of the rectangles and through these points draw a smooth curve F which is termed an approximate *frequency curve*⁴ for our data. Since the area under any section of the curve F is approximately that under the corresponding part of the histogram, it is clear that F also possesses the probability property which was noted in the case of the histogram.

Assume, now, that we have conducted a new firing program in which many more projectiles were shot than before. Since better classification would be possible because of the added data let us

³ Introduction to the Theory of Statistics, G. U. Yule, page 84.

⁴ Introduction to the Theory of Statistics, G. U. Yule, page 87.

make smaller zones in figure 1 as indicated by the dotted lines. Let the results of the firing be shown by the numbers in rectangles in each of the new zones. For these new data the histogram and frequency curve F_1 are given in figure 3. The total area under F_1 is different from that under F but the probability property holds for F_1 in exactly the same form as it did for F .

Let us draw curve F anew with the vertical distances of each point from the horizontal axis reduced by such a factor that the area under the new form F' of the curve will be 1. Suppose that the same procedure were followed in the case of the curve F_1 giving a new curve F'_1 , the area under which would also be 1. In the present problem the curves F' and F'_1 are so nearly identical that, as a consequence, in figure 4 only one curve is given which may at will be considered as F' or as F'_1 . In figure 4, after the reduction of the verticals mentioned above was made, the vertical scale was magnified in order to give a clear figure. Since in passing from figure 2 or figure 3 to figure 4 all parts of the area under F or F_1 were changed proportionately, it follows that, since the area under F' or F'_1 is 1, the probability property holds for F' or F'_1 in the following simplified form:

On the basis of the data from which F or F_1 was obtained, the probability of a shot falling between A and B in figure 1 equals the area under the curve F' or F'_1 in figure 4 between the verticals through A and B .

Let us call F' and F'_1 *approximate probability curves*.

If we made more and more firings, classified the results in zones of smaller and smaller width, drew the corresponding frequency curves and, from them, the approximate probability curves, the latter would approach a definite limiting form. As a result of certain theoretical conclusions to be quoted in Section 2 it can be stated that this limiting form would be a *normal probability curve*; that is, a curve whose equation in coördinates (x, y) , as indicated in figure 4, would be given by the equation

$$y = \frac{h}{\sqrt{\pi}} e^{-h^2 x^2} \quad (\pi = 3.1416; e = 2.7183) \quad (1)$$

for some value of the positive constant h . In the present ideal firing problem the value of h is approximately 0.48. A different normal curve would be obtained for every type of gun which might be used in the firing, or, in other words, a different value of h would be determined for which formula (1) would represent the probability curve

for the gun in question. For a given gun, the probability curve would change if the distance of T from the gun were changed.

In actual firings at the Artillery Proving Grounds of our army, when it is necessary to determine the probability curve corresponding to a given gun when fired at a target T at a given distance, the mode of procedure differs from that outlined above. From the firing results as given, for example, in figure 1, an approximate value of h is immediately found by means of a formula to be given in Section 3 and the probability curve is then constructed by means of formula (1) above.

As a second example of a statistical nature consider the following analysis of the weights of men in England, Ireland, Scotland and Wales. This problem⁵ is cited by G. U. Yule who obtained the data from the Final Report of the Anthropometric Committee to the British Association (1883), page 256. In the table below, the first column gives the lower ends of the 10-pound weight intervals in which the weights of the men were classified. The second column gives the number of men whose weight fell in the corresponding interval.

WEIGHT	NUMBER OF MEN
<i>pounds</i>	
90	2
100	34
110	152
120	390
130	867
140	1623
150	1559
160	1326
170	787
180	476
190	263
200	107
210	85
220	41
230	16
240	11
250	8
260	1
270	0
280	1

⁵ Introduction to the Theory of Statistics, G. U. Yule, page 95.

The mean, or average, weight of the 7749 men examined is found to be 157 pounds.

On treating these data by a method similar to that used in the artillery example, we obtain a histogram and a frequency curve G as given in figure 5. The heavy vertical line in this figure goes through 157 on the horizontal axis, the point corresponding to the mean of the weights measured. Consequently, the area under G to the right of this vertical equals the area to the left. On the basis of the data used in this example, the probability that a man selected at random would have a weight between A and B pounds equals *the area under G , between vertical lines in figure 5 through A and B on the horizontal axis, divided by the total area under G .*

If an approximate probability curve G' should be obtained from G as was done in figure 4 of the previous example, G' would retain the property, possessed by curve G , of being unsymmetrical with respect to the vertical line through 157 on the horizontal scale and unsymmetrical with respect to the vertical line through the highest point of the curve. Hence, the curve G' could not be approximately represented by a normal probability curve because, in such a curve, the area beneath it is divided into two equal parts by the vertical line through the highest point. Nevertheless, if the data for the present example were progressively increased and were then classified more finely, and if the corresponding approximate probability curves were drawn, they would approach a limiting curve which, however, would not be a normal probability curve. In spite of this fact, the limiting curve would possess the probability property and would be very useful in any further discussion of the problem.

To almost every statistical problem of the type considered here there corresponds some sort of a limiting probability curve. In a great majority of cases, the problems are of such a type that it is impossible to state at the beginning, as was done in the artillery problem, that the probability curve will be normal. Hence, the only proper and safe mode of procedure in a statistical investigation is to construct the frequency curve and then to determine by inspection whether or not it could possibly lead to a normal probability curve. If it is found experimentally or by theoretical reasoning that the probability curve for the problem is normal, the properties of the latter may then be used in the further discussion of the example.

2. *The normal curve as a curve of errors.* If a given quantity is being measured, errors naturally enter in the use of the scales of the measuring instruments. For example, in five measurements of the length of a bar we might obtain the following results: 16.16, 16.18, 16.17, 16.19, and 16.15 inches. The arithmetic mean, or average, of these results is 16.17 inches. The *deviations* of the measurements from the mean are: -0.01 , $+0.01$, 0.0 , $+0.02$ and -0.02 inch. Suppose that we have made a long series of different measurements of the size of some given magnitude and let us tabulate the deviations of the individual measurements from the mean of all the measurements. Then, as in Section 1, from these tabulated data we could obtain a frequency curve F showing the frequency of occurrence of deviations of various sizes. From F in turn we could obtain an approximate probability curve F' . In regard to this last curve we can state the following theorem:⁶

Theorem. Suppose that a given magnitude is being measured and assume that errors in the individual measurements are wholly due to accidental causes. Then, as the number of measurements becomes infinitely large, the approximate probability curve F' , showing the probability of occurrence of deviations of various sizes, approaches a normal probability curve.

The term *accidental* is applied to errors which are due to irregular causes which operate to increase observations as often and in as great a degree as they tend to decrease them, and whose effect upon an observation is independent of the individual character of that observation. The practical significance of the words *infinitely large* is that we should not expect an approximate probability curve for a problem to correspond very closely to a normal curve unless the number of entries in our data was very large.

The theorem above is stated in terms of the word *measurement* but the result remains true for any set of data which possesses the same characteristics as a set of measurements subject only to accidental errors of observation. Therefore it follows from the properties of accidental errors that, if the theorem holds for a set of numerical data, the deviations of the individual entries from the mean of the whole set of data must satisfy the following conditions:

⁶ For proof of this theorem see Method of Least Squares, D. P. Bartlett, Chapter 1, or Calcul des Probabilités, H. Poincaré, Chapter X. For logical objections to the proof of the theorem see Poincaré, loc. cit., page 173.

(a) Positive deviations are as frequent as negative deviations of the same magnitude.

(b) The probability of the occurrence of large deviations is very small and that of small deviations is relatively large. It is easily seen that the frequency curve for such a set of data must necessarily be of the symmetrical type met in the first example of Section 1. The maximum of such a curve comes directly above the point on the horizontal scale corresponding to the mean of the whole set of data.⁷

Consider the artillery example of Section 1 in its relation to the theorem. Deviations of the points of fall of the projectiles from the target T were due wholly to accidental errors in the manufacture of the projectiles and of the powder charges. Hence, deviations of the points of fall, which make up the data considered in Section 1, must have the same nature as the accidental errors of manufacture by which they were caused. Consequently the theorem applies and therefore the probability curve for the problem must be normal.

Consider the second example of Section 1. In this case the frequency curve was found to be unsymmetrical. Hence, conditions (a) and (b) above cannot hold and therefore the probability curve for the problem is not normal.

It must not be inferred that the conditions of the theorem are the only ones under which it can be established that a normal curve results. The probability curve can be derived from many different points of view⁸ and only one has been presented in this section.

3. *Properties of the normal curve.* The graph in figure 4 shows the form of the normal probability curve for the special value $h = 0.48$. Consider for a moment the five measurements of the length of a bar given in Section 2. The deviations from the mean were -0.01 , $+0.01$, 0.0 , $+0.02$, and -0.02 inch. If we add these deviations, making all the signs positive, we obtain 0.06 inch. The *average deviation*—in notation, the A. D.—of the observations is defined as $0.06 \div 5 = 0.012$. For a general set of data we similarly define the A. D. of the data as *the sum of all the deviations taken with a positive sign divided by the number of entries making up the data.*

⁷ A very complete study of such frequency curves is found in Yule's *Introduction to the Theory of Statistics*, Chapter VI.

⁸ See *Calcul des Probabilités*, Poincaré, Chapter X, and *Introduction to the Theory of Statistics*, Yule, Chapter XV.

The A. D. of a set of data is very easy to compute⁹ and serves as a basis for determining other properties of the data.

Assume that we are dealing with a group of data for which the probability curve is normal (suppose its graph is given by figure 4). Then, it can be easily proved¹⁰ that the value of h in formula (1) which gives the normal curve for the problem is related to the A. D. of the data by the equation

$$h = \frac{1}{\sqrt{\pi}(\text{A.D.})}.$$

In the artillery example of Section 1 the A. D. for the group of data in circles in the large zones is found to be 1.194 which gives $h = 0.47+$. For the group of data in rectangles, the average deviation is 1.164 and the corresponding value of h is $0.48+$. In a statistical problem where the probability curve is desired, the most simple procedure usually is to compute the A. D. of the data and from it the value of h , after which the probability curve can be obtained by formula (1). This method replaces the more involved steps illustrated in Section 1.

In regard to figure 4 it was stated that the probability of a shot falling between A and B in figure 1 was equal to the area under the probability curve lying between vertical lines erected at points A and B on the horizontal axis in figure 4. For a general set of data, we may state that the probability of a deviation having a value lying between A and B is equal to the area under the probability curve between the verticals at A and B . The area under the whole curve is 1. Therefore, there are points, $+R$ and $-R$, on the horizontal axis such that the area under the probability curve between the verticals erected at these points is $\frac{1}{2}$. In other words, R is such a number that the chances are even that a deviation will exceed R in numerical value. This value R for a set of data is called the *probable error* of the data, and it can be proved¹¹ that it is related to the A. D. by the equation $R = 0.8453 (\text{A. D.})$. In the artillery example, the two sets of data give respectively, $R = 1.01$, and $R = 0.99$. It can be proved that there is less than one chance in 100 of a deviation being greater than $4R$ in numerical value. Another quantity,

⁹ See Introduction to the Theory of Statistics, Yule, Chapter VIII.

¹⁰ Method of Least Squares, Bartlett, Chapter III.

¹¹ Method of Least Squares, Bartlett, Chapter III.

called the mean error, M , or *standard deviation*, is of importance in its relation to the normal curve. This quantity¹² satisfies the equation

$$M = \frac{1}{h \sqrt{2}} = 1.4826 R$$

It is beyond the scope of this paper to consider the applications of the probability curve, the mean square deviation and the probable error in statistics and the reader who desires a brief discussion of these should consult a paper¹³ by E. V. Huntington on "Mathematics and Statistics."

¹² Introduction to the Theory of Statistics, Yule, Chapter XV, where the subject is discussed in great detail.

¹³ American Mathematical Monthly, Vol. 26 (1919), page 421.